Improving Loudspeaker Characteristics in a Low Power Environment

JONAS ANDREASSON & LOVE OLSSON
MASTER´S THESIS
DEPARTMENT OF ELECTRICAL AND INFORMATION TECHNOLOGY
FACULTY OF ENGINEERING | LTH | LUND UNIVERSITY

Improving Loudspeaker Characteristics in a Low Power Environment

Jonas Andreasson & Love Olsson
{jo7334an-s, lo0016ol-s}@student.lu.se

Department of Electrical and Information Technology
Lund University

Supervisor: Kaan Bür

Examiner: Maria Kihl

June 14, 2024

© 2024
Printed in Sweden
Tryckeriet i E-huset, Lund

Abstract

This master's thesis explores the improvement of loudspeaker characteristics by using digital signal processing. The main aim of the thesis is to improve the overall audio quality with regards to limited processing power and power supply. Improving the audio quality is important since the vast expansion of the usage of loudspeakers has led to demands on cost, size and design choices all leading to different loudspeaker characteristics. A software algorithm capable of running in real-time on the loudspeaker is used to achieve the aim. The thesis includes a literature study as well as a practical implementation with tests to find out which algorithm performs the best. The result of the literature study concluded that a virtual bass enhancement algorithm, as well as a dynamic equalizer algorithm, are the best alternatives to enhance the audio quality of the network loudspeakers. From the many different algorithms found in the literature study, the virtual bass enhancement algorithm, arc-tangent square root, was determined to give the best result by both scoring the highest of the algorithms tested in a listening test and showing good performance in performance tests. The simplicity of the algorithm together with the improvement it offers, leads to it being a good option for any network speaker. Future improvements on the algorithms should focus on finding a better way of reducing noise on the input signal before processing it with the algorithms.

Popular Science Summary

In today's society, speakers play an integral part of everyday life. Due to speakers existing everywhere including your phone, at the train station, and in the elevator, there is value in researching how to improve the audio quality for all loudspeakers. The problems that occur when design choices, such as form factor or economical considerations, impair the ability of the loudspeaker to produce bass has historically been solved using a sub-woofer, a specialized bass speaker. However, with the demand for smaller speakers ever increasing, using a sub-woofer to enhance the bass is not feasible because they need to be bigger than a regular speaker. This project focuses on how the audio quality can be improved on a network speaker with the help of software.

By using digital signal processing the quality can be improved with a few methods. In this project two of these methods were used. One method focuses on making the bass more pronounced. The other method focuses on tricking the brain that there is more bass than there actually is.

A simple way to make the bass more pronounced is with the use of digital filters, in this case a low-shelf filter, which will boost a certain range of frequencies, while leaving the rest unmodified. However, the signal can only reach a certain level before information is lost. The space left between the peaks and the maximum level is commonly called headroom. Processing the signal with a low-shelf filter will decrease the amount of available headroom. By dynamically changing the amount boosted with regards to the amount of available headroom a more pronounced bass can be achieved, while not disturbing the rest of the audio content. This approach is a so called dynamic equalizer (EQ).

Tricking the brain into thinking there is more bass can be done by using a psychoacoustic phenomenon called "Missing Fundamental". This phenomena occurs because pure tones are not natural. Natural tones consists of a fundamental tone and the fundamental's harmonic series, also known as overtone series. In the overtone series the fundamental is the strongest and the overtones are slowly decreasing in strength. The "Missing Fundamental"-phenomenon makes it so, if a tone's harmonic series can be heard then the fundamental can be heard as well, even if the fundamental does not exist in the signal. Because of that, if the harmonic series could be generated the brain would think the bass was more pronounced than it actually is. With the help of a non-linear device (NLD), the harmonic series of a signal can be generated.

A basic way to create a better perceived bass with the help of an NLD is by simply allowing the bass content of a signal to be processed by an NLD. This was implemented with the help of the NLD arc-tangent square root (ATSR). The ATSR takes an input signal and applies a mathematical function to calculate the resulting signal.

Another more advanced way to improve the perceived bass is to generate the harmonics with regard to different instruments instead of the signal as a whole. This method consists of trying to "demix" the signal into multiple signals, each

representing different instruments and then applying the NLD to the different signals. In the implementation that was done the signal is not "demixed" and instead it is treated as one single instrument.

The two implementations mentioned earlier together with a dynamic EQ were evaluated by participants of a listening test. In the listening tests, the participants were instructed to rate each algorithm for six different music clips and six different announcements. The results showed that the basic ATSR implementation improves the perceived audio quality for music compared to the unaltered reference signal. While the dynamic EQ algorithm shows improvement on some of the music clips compared to the reference, it is not rated as high as the basic ATSR implementation. The music demixing implementation did not show any significant improvement to the audio quality. Neither of the implementations showed any positive change in the audio quality on the announcements. The three implementations were also tested with a performance test measuring the processing power required to run the implementation. The performance test showed that the dynamic EQ had the best performance, closely followed by the basic ATSR implementation. The music demixing implementation had the worst performance on the performance test.

Finally the conclusion of this project is that the audio quality of any network loudspeaker could be improved by implementing the basic ATSR implementation presented. For scheduled announcements, the algorithm could be turned off momentarily. Therefore, the ATSR should be implemented for any network loudspeaker.

Further research should focus on implementing a noise reduction filter. By reducing the noise in a better and more effective way all of the algorithms tested in this project would be improved and the audio quality would be improved even more. Additionally, testing more types of speakers than the ones in this thesis would be beneficial to achieve a wider understanding of the algorithms.

Acknowledgements

We would like to thank our supervisor Kaan Bür for his help and guidance during this masters thesis. We would also like to thank our examiner Maria Kihl for her constructive comments and patience.

Additionally, we want to extend our thanks to the partnered company and our on-site supervisor Erik Fröbrant for helping us with our problems during this masters thesis. This thanks is also extended to those employees who helped us in our project and participated in our listening test.

I, Jonas, would like to thank my partner for her input on certain sections and validating that any motivations in the report were reasonable.

Finally, I, Love, would like to thank my partner Rebecca for supporting me during this time with both English and encouragement, as well as my dog Stevie for cheering me up throughout this masters thesis.

Abbreviations

ATSR	Arc Tangent Square Root
BPF	Band-Pass Filter
CMOS	Complementary Metal-Oxide-Semiconductor
CPU	Central Processing Unit
CSV	Comma Seperated Value
DSP	Digital Signal Processing
EQ	Equalization
EXP	Exponential
FA	Frequency Analysis
FFT	Fast Fourier Transform
HPF	High-Pass Filter
IIR	Infinite Impulse Response
LSF	Low Shelf Filter
LPF	Low-Pass Filter
NLD	Non-linear Device
NTANH	Normalized Hyperbolic Tangent
PoE	Power over Ethernet
SSH	Secure Shell
STFT	Short Time Fourier Transform
VB	Virtual Bass
VBE	Virtual Bass Enhancement

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In today's society, loudspeakers play an integral part in everyday life. Used for such things as announcements when waiting for a train, making sure that the stand up comedian is being heard clearly on their show, and for listening to music and entertainment. These are some of the applications of the loudspeaker and a common denominator is that the best possible audio quality is desired. But due to design choices such as form factor, economical factors, and looks, this is not always possible. Therefore, there is value in the preprocessing of audio to make the audio quality better for every speaker. In extension, finding what type of process that can be applied to gain maximum performance in audio quality is of much interest. Since loudspeakers get smaller and smaller to be able to fit them everywhere, the ability of a speaker is impaired. Due to these size restrictions, the speaker will not be able to reproduce the desired frequency range. With the increased digitalization, the desire for connectivity has also increased. This has lead to many companies using Power over Ethernet to connect their loudspeakers. Due to the limitations of Power over Ethernet, another challenge is introduced in the form of power limitations, further limiting the possible solutions. This thesis' aim is therefore to improve on these problems by using digital signal processing with regard to a more restrictive power supply.

1.1 Thesis Aim

The aim of the thesis is to investigate and answer the following points:

- Investigate what kind of digital signal processing can be applied to enhance perceived audio quality on any speaker.
- Investigate how the signal processing can be carried out in an optimal way under various conditions (e.g., different volume levels, limited frequency range).
- How quickly can the signal processing be adjusted when the audio content changes, such as altered signal level, without causing artefacts?

The reasoning behind investigating what kind of digital signal processing can be applied, is to find a useful method where the perceived audio quality can be enhanced the most. Due to the subjectivity of audio quality there are many different ways to improve upon it. The signal levels and frequencies also have to be taken into account. This combined with the last point need to be considered due to artefacts affecting the audio quality negatively. To minimize these artefacts, the audio can not be boosted to an extreme degree.

1.2 Delimitations

This thesis was carried out during a five month period, emphasizing the importance of specifying the scope of the project. To ensure the project's scope not growing too big, strategic choices were made when selecting the focus of the project. The limitations of this study include:

- To only use network loudspeakers produced by the partnered company, even though the project aims to be applicable to all loudspeakers.
- To only test music and spoken voice for the listening tests.
- To limit the amount of algorithms implemented, and to only test three algorithms in the performance and listening tests.
- To limit the performance metrics measured to only Central Processing Unit(CPU) usage.

The decision to only use network loudspeakers from one producer was based on the master thesis being done in collaboration with a local company. Therefore, the availability for their speakers was significantly higher than that of other manufacturers. This also guarantees the compatibility between different speakers and the software. The thesis was limited to only test certain metrics as well. In the case of the listening test, only music and voice were tested, and not other audios such as sine sweeps. This limitation was applied because the algorithms operate in different ways rendering tests, like sine sweep, less effective for measuring quality of the audio. For the performance metrics only CPU usage was tested since it is a good indicator if an implementation can be used in practice when there are other processes running. Finally, the amount of algorithms implemented and tested were limited due to different reasons. In the case of the amount of algorithms implemented, the thesis is time-limited and focusing on a few implementations will yield a better result. For the ones that were tested in the listening and performance tests, the amount of algorithms were limited to have a more suitable amount of implementations for the test subjects to rate.

1.3 Approach

In this thesis the scientific method was used. The scientific method consists of multiple steps to ensure a successful thesis. The first step in the scientific method is to ask a question that you then try to answer. Then when a clear problem has been established, the topic is researched to gain knowledge about the issue. With enough research, a hypothesis can be presented. This hypothesis is then tested and evaluated. After the tests, the data can be analyzed and a conclusion can be drawn.

The aims for this thesis were established early to make sure there was a clear goal in mind. This was done by asking the questions seen above. The first step also included creating a plan covering all of the parts of the thesis, clearly defining at what week a certain part should be completed. When the planning was finished, the literature study, or research, could commence and information for the theoretical background as well as algorithms were found. The third step, and hypothesis, was the suggestion of what algorithms to implement and the implementation of those algorithms. This step was done after all the information had been gathered from the literature study. After the algorithms had been implemented and decided upon, listening tests were held to gather data on what algorithm people thought sounded the best. These results were then gathered, presented and discussed in the final step, drawing conclusions on if the goals were reached.

1.4 Disposition

Chapter 1 consists of the introduction and introduces the background, motivation, main goals and challenges in the thesis.

Chapter 2 is the result from the literature study. Here the theory behind audio, loudspeakers and approaches are explored more in depth.

In chapter 3 the implementation of the algorithms is introduced. The argumentation for what to implement and why is also presented.

Chapter 4 describes the tests that were carried out. In addition, the results of the listening tests and performance measurements are presented together with the evaluation of them.

Finally, in the closing chapter 5, the culmination of the study is presented with suggestions for future research.

Chapter 2
Theory

In recent times, there has been a interest in smaller loudspeakers, primarily for the use in mobile phones, hearing aids, and "in-ear" headphones [1], [2]. Ideally, these small, or even tiny, loudspeakers should have a flat frequency response with low power usage and high energy efficiency. The demand for better audio quality and a higher audio level has also increased [3]. One major issue with smaller loudspeakers is the low-frequency response. The speaker is designed to vibrate the air in-front of the speaker. When frequency decreases, the air load against the speaker decreases as well [4]. With less air pressed against the speaker, the speaker will affect less air, making the sound fainter. This is the case until the frequency matches that of the speaker's mechanical resonance. Below the mechanical resonance, the problem is not primarily the lack of air, but the speaker's inability to move it. The stiffness of the membrane requires more power to bend, compared to higher frequencies, resulting in a lower volume. Reducing the size of these conventional loudspeakers seems to also reduce the frequency range of the speakers [5].

2.1 Type of Loudspeakers

There are primarily four types of speakers: electrodynamic, electrostatic, piezoelectric, and thermoacoustic. They differ on what mechanic they use for actuation and they all have different advantages and disadvantages.

2.1.1 Electrodynamic

Electrodynamic loudspeakers use electromagnetism for actuation and are the most common variety of loudspeakers [1]. An electrodynamic loudspeaker works by regulating an electromagnetic field causing a coil to move. The coil is attached to a diaphragm which in turn vibrates the air; producing sound. This technique possesses a high power density with a low driving voltage and a linear response [1], [6]. However, this type of actuation requires a permanent magnet, which could restrict the implementation in smaller loudspeakers [1].

2.1.2 Electrostatic

Electrostatic loudspeakers apply electrostatic charges between two electrodes to make one of them move [7]. This technique has been able to stay prominent on the micro- and nano-scale thanks to the size only being restricted to how close two electrodes can be positioned, which is commonly referred to as the "electrode gap" [8]. One advantage with this type of actuator is a theoretical smaller travel range however, in reality, this is restricted by the pull-in effect which could lead to device failure [8]. Another advantage is that the electrostatic speakers are compatible for complementary metal-oxide-semiconductor(CMOS)-integration [8], [9].

2.1.3 Piezoelectric

Piezoelectric loudspeakers utilize piezoelectric elements to directly push the output mechanism. This leads to the piezoelectric drivers having high accuracy, large output load, and can be designed with a small form factor [10]. In addition,

piezoelectric speakers also have no electromagnetic interference as well as do not create a magnetic field, making it well suited for various applications.

2.1.4 Thermoacoustic

A thermoacoustic speaker does not use any actuators, unlike the previously mentioned types. It works by heating up a material through the use of an electric current, which causes it to oscillate, meaning it requires no moving parts to produce sound [11], [12]. A popular material to use as the oscillator is graphene [11], [12], [13], [14]. These graphene speakers can achieve great sound pressure level with low total harmonic distortion, except in low frequencies [12].

2.2 Audio Network Amplifiers and Speakers

There are multiple different audio network amplifiers, and these amplifiers are designed to work over Ethernet for streaming and power supply. There are three network amplifiers and speakers tested in this project, where all of the amplifiers are designed to work with Power over Ethernet (PoE) and to operate at max 12.9 W and 25 W. The first amplifier, considered the least powerful in terms of processing capability, features an i.MX 6ULL chip with 256 MB of random-access memory (RAM) [15]. The second amplifier, considered mid-range in processing power, includes an i.MX 6SoloX chip with 512 MB of RAM [16]. The third and most powerful network amplifier, is equipped with an i.MX 8M Nano chip and 1024 MB of RAM [17]. The first and third amplifiers also have an electrodynamic speaker in the same enclosure while the middle amplifier can be used together with all passive loudspeakers to transform them to a network speaker [15], [16], [17].

2.3 System Description

For this master's thesis, the partnered company's network amplifiers and speakers will be used. These speakers are designed for PoE, leading to the management of power consumption being crucial since there is only a set amount of power available to the speaker and amplifier. An issue with these network speakers, like other smaller speakers, is their frequency response, specifically in the bass registers. This could be resolved in three ways:

- Increase the available power
  Increasing the available power makes it possible to boost the frequencies that are hard for the speaker to produce to a more reasonable volume.
- Increase the size of the speaker
  Increasing the size of the speaker makes the effect described in previous sections less potent.
- A smart algorithm that improves the perceived sound.
  Instead of trying to change the actual frequency response, an algorithm that improves the perceived audio rather than the actual audio could be preferable.

It can be argued that there is only one way to achieve a better perceived frequency response, since increasing the available power is not possible when following PoE standards and making the speakers bigger would restrict any creative freedom of the designers. Therefore, the best way, without physically changing the loudspeaker, to improve the loudspeakers' sound is by using a digital method to enhance the audio. The method is still restricted to the processing power of the chips integrated in the speakers, as well as the maximum power available through PoE. Furthermore, it is preferable if the method does not induce a high latency for the rest of the system, i.e. introducing delay between a user's input and the speaker's response. Therefore, an efficient algorithm to improve the perceived low frequencies is to be implemented and tested. The system for where these algorithms should be implemented can be seen in Fig. 2.1. It is possible to see that

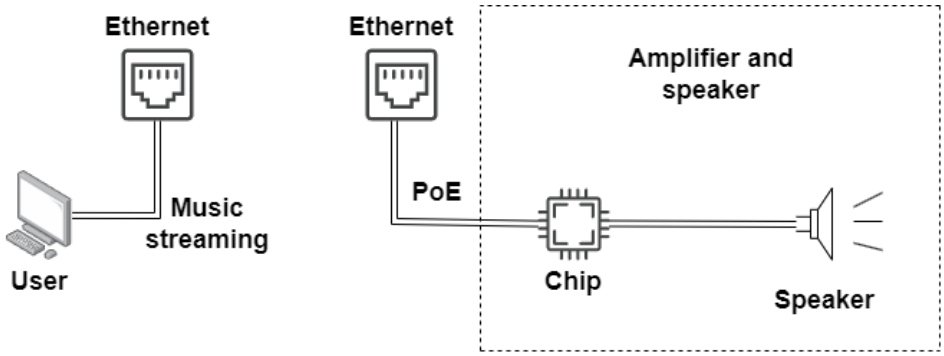

Figure 2.1: Figure showing the system setup for this master's thesis.

a user streams audio over Ethernet to the speaker. The speaker then receives the audio and power through Ethernet. This is where the algorithm is supposed to work in real time at the amplifier part of the network speaker. The algorithm implementation will be tested on the chip of the network amplifier. For simplicity in the listening tests the audio clips will be preprocessed.

2.4 Audio Signals

An audio signal can be represented to a high degree of accuracy as a sum of several sinusoidal waves of different amplitudes and frequencies, all possibly varying in time. For a clip of music or a person speaking the amplitude is the maximum absolute value of the signals displacement from zero, meaning how loud or quiet the signal is. The frequency of the signal is called the pitch. Lower amplitude means lower volume and lower frequency means lower pitch. This analog signal is hard to work with and needs to be converted to a digital signal before it can be processed with digital signal processing [18]. In this project the signals are already in a digital state.

2.5 Power over Ethernet

PoE is a technology used for supplying devices with power in parallel to data through ethernet. PoE (Type 1) is limited to supplying 12.95 W of power and has since 2003 been a standard from IEEE [19]. The newest version of PoE (Type 4) can supply up to 71.3 W of power [20].
Some benefits of PoE are highlighted in [21] and include:

1. Time and cost savings
   Since there is just one cable both time and money will be saved on installation and maintenance.
2. Flexibility
   PoE makes it possible for multiple devices to connect to it such as cameras and network speakers.
3. Safety
   PoE is designed to in a smart way protect the network device from overload, incorrect installation, and underpowering.
4. Reliability
   PoE comes from one source instead of using multiple individual wall adapters. It can also be assisted by an uninterruptible power source meaning it can keep supplying power in the case of any outages, as well as easily reset or disable devices.
5. Scalability
   Connecting new devices will be easy and straightforward.
6. Enhanced productivity
   PoE increases productivity since it uses the same cable for power and data, helping the devices to effectively collect data.

2.6 Enhancing Audio Quality

There are several ways to improve the perceived audio quality of a signal. The following section will highlight some methods to achieve an improved perceived audio quality.

2.6.1 Equalization

The purpose of audio equalization (EQ) is to change the volume of certain frequency ranges, most commonly referred to as bass, mid, and treble being the lower, mid-range and higher frequencies respectively. What sounds best is often highly subjective and these controls on speakers or software makes it possible for any person to dial in what they think sounds the best. Different frequencies, and their respective loudness, are however not preserved relative to each other, and therefore the balance of frequencies in the sound varies depending on the listening levels. This is an effect that we have all heard when the volume of a recording is turned down and the bass and treble components sound quieter relative to the

midrange frequencies, and the sound becomes “duller” and “thinner" which most people do not prefer [22]. One way to improve the bass of the speaker is to apply a low-shelf filter (LSF) to boost the bass. This comes with several problems, such as requiring more power which is not guaranteed to be accessible, and potential bleed over into the easily reproducible ranges causing audio peaks as well as damaging the speaker.

Dynamic Equalization

Another way of attempting to achieve a better sounding bass is to use a technique called dynamic EQ. The idea of dynamic EQ is to dynamically change the filtering and modification of the input signal, depending on how the audio signal behaves. Headroom is the space between a signal peak and the maximum possible amplitude before clipping occurs. For example if there is more headroom in the signal than usual we can boost the bass without it clipping, or if there is less headroom it is possible to attenuate the mid and higher frequencies to make the bass appear louder. There has also been research that focus on boosting the bass without boosting noise when there is a noise source nearby, i.e engine noise in a car cabin [23]. Dynamic EQ has also been implemented to achieve equal loudness over the frequency spectrum [24]. This means that any frequency should appear to be as loud as any other frequency.

2.6.2 Virtual Pitch and Virtual Bass Enhancement

Normally, an amplifier aims to have a linear relation between the input and the output [25]. However, this is not the case when using a non-linear device (NLD). NLDs introduces a non-linear relation between the input and output. A downside in electronics is the fact that NLDs introduce harmonic distortion. However, introducing the harmonics can be beneficial when dealing with audio processing.

The missing fundamental is a pyschoacoustic effect that allows the listener to hear the fundamental tone when only given the harmonic series [26]. The idea that harmonic series enhances the fundamental tone was first theorized by Seeback [27], and can be seen in Fig 2.2. In the figure the gray arrow shows that by having the harmonics of the fundamental frequency present in the audio signal, the perceived amplitude of the fundamental can increase and appear to be louder than it is.

Virtual pitch is the exploitation of "the missing fundamental". This allows for the "reintroduction" of fundamental tones that a speaker is not able to properly produce. For low frequency, bass tones, this is referred to as virtual bass (VB) and when applied on speakers using digital signal processing (DSP) it is called virtual bass enhancement (VBE). There are several works that have used this mechanic to enhance the audio quality. This has been achieved either through a pure NLD implementation, a pure phase vocoder (PV) implementation, or hybrid approach. A PV is a technique when processing audio signals that uses a Short Time Fourier Transform (STFT) to study a signal in the frequency domain [28], [29]. The PV

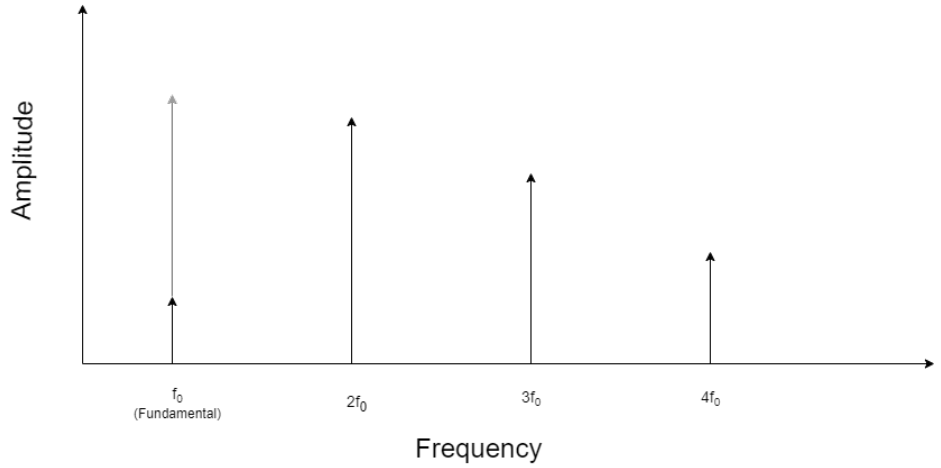

Figure 2.2: A visualization of the missing fundamental concept.

allows for better harmonic control [30]. A major downside with a PV is transient smearing, resulting in the loss of percussive elements [28], [30].

In [31] multiple (Arc-tangent Square Root (ATSR), Normalized Hyperbolic Tangent (NTANH), and Exponential (EXP)) NLDs were implemented in hardware for real-time audio processing in an attempt to increase the experienced bass. Furthermore, it was shown that all three NLDs performed better in a subjective listening test, showing a higher level of bass in a high passed signal compared to an anchor. This is echoed in [32] for speakers that have a higher cut-off frequency.

As mentioned earlier, another approach is by using a PV to improve the perceived bass. In [33] it is shown that a PV can improve the bass impression, however the audio quality suffers. It is also shown that a PV is significantly worse compared to an NLD. This statement is also reflected in [30] where it can be seen that an NLD performs significantly better in overall sound quality compared to a PV. Furthermore, both [30] and [33] shows that a hybrid solution of PV and NLD scores higher in bass impression as well as overall sound quality compared to a pure NLD or pure PV solution. It should also be stated that the hybrid solutions show an improvement compared to an anchor in the case of [30] and a reference in [33].

2.7 Proposed Solutions

The following methods and algorithms are deemed sufficient to make progress on solving the main problem of this thesis. This section will highlight the reasoning as to why, as well as information that was directly taken from other papers.

2.7.1 Non-Linear Device

Several papers [30], [31], [33] have highlighted the viability of using an NLD to improve the perceived audio quality by using virtual bass enhancement as detailed above. According to a mathematical analysis of different NLDs and subjective listening test highlighted in [34], one of the following NLDs are preferable due to what is described as "good" related to its ability to enhance the bass.

$$\mathrm{EXP}(x) = \frac{e - e^{1-x}}{e - 1} \tag{2.1}$$

$$\mathrm{NTANH}(x) = \frac{(e^{x} - e^{-x})(e + e^{-1})}{(e^{x} + e^{-x})(e - e^{-1})} \tag{2.2}$$

$$\mathrm{ATSR}(x) = 2.5 \arctan(0.9x) + 2.5\sqrt{1 - (0.9x)^2} - 2.5 \tag{2.3}$$

Where x is limited between -1 and 1 and corresponds to the amplitude of a signal. The following equation is a variation of ATSR found in [31] with the same variable.

$$\mathrm{ATSR}_{\mathrm{var}}(x) = 2.5 \arctan(0.9x) + 2.5\sqrt{100 - (0.9x)^2} - 25 \tag{2.4}$$

In [31], NTANH, EXP, and $\mathrm{ATSR}_{\mathrm{var}}$ are shown to have a significantly higher perceived bass after removing the actual bass content compared to a signal which has only had its bass removed. With [34] showing that ATSR scored the highest in a subjective listening test rating the amount of perceived bass. In the listening test the subjects were asked to rate different NLDs compared to a reference on a scale of -3(less bass) to +3(more bass) on three different songs, with ATSR scoring an average mean score, across all three songs, of ~ 2.73. The listening test also included a hidden reference which scored an average mean score of ~ -0.01. Showing that ATSR could be a solution to this master thesis' problem. A simple NLD implementation is highlighted in [34] where the input is split in two by applying a low-pass filter(LPF) and high-pass filter(HPF) respectively. The low-passed signal is then processed with an NLD before summing the two signals to from the output.

2.7.2 Dynamic Equalizer

A dynamic EQ shows a possible bass enhancing effect with low distortion in [35]. Since it was only testing the bass enhancement and the distortion it can be argued that the overall audio quality could improve compared to an unmodified signal.

2.7.3 Music Demixing with Non-linear Device

There have been attempts to separate different parts of an audio signal to process only specific parts of the signal or different parts differently. One way is described above by only applying the NLD to a low-passed signal. Another way is described in [36], where it is suggested to split the audio-signal with regards to the different instruments or groups of instruments which is done by applying a Music Demixing Model. The solution described in [36] then processes each of the individual music

tracks by applying an LPF and subtracting this from the original signal. The low-passed tracks are then normalized before processing it with an NLD and weighing the NLDs output. Finally, all of the tracks are summed with the modified original signal. This approach resulted in a higher basic audio quality compared to a reference and a significant amount of bass enhancement in a subjective listening test.

2.7.4 Hybrid Phase Vocoder and Non-linear Device

In [30] it was shown that a hybrid PV/NLD solution could outperform a regular NLD in overall sound quality as well as the amount of bass. The hybrid solution works by separating the signal in a low-frequency content signal and a high-frequency content signal. The low passed signal was then processed by creating a beat spectrum by using a STFT then finding the peaks and separating them from the original signal. This was done by applying a notch filter around the found peak frequencies and subtracting the result from the original signal. The separated peaks are then converted back to time-domain and processed with an NLD. Finally the low-passed signal is boosted to better match the gain of the high-passed signal. Finally it is combined with the high-passed signal to form the output signal. In [37] a hybrid method is shown that significantly increases bass quality with minimal unwanted distortion compared to a regular NLD. This method works, like many other methods previously mentioned, by splitting the signal in a low-frequency component and a high-frequency component. Then processing the low-frequency component before summing it with the unmodified high-frequency component to form the output signal. The processing works by downsampling the signal and preforming a STFT. The output of the STFT is then processed with a "Harmonic and Percussive Component Separation"-element, splitting the percussive and harmonic parts of the signal in two. The harmonic part is then analyzed to find the peak before converting this back to the time domain and processing it with an NLD. This NLD processed harmonic component is then combined with the percussive element which has been band-boosted. Finally this combined signal is processed with a band-pass filter(BPF) and up-scaled to finally be summed with the high-frequency component as mentioned before.

2.8 Listening Tests

As previously stated, audio quality is highly subjective and needs to be quantified in some way for it to be reliable data. For listening tests there are three main problems to keep in mind while designing these test, and these are:

1. Reproducible at different times and places, with different listeners.
2. Reflect only the audible characteristics of the system under examination.
3. Reveal the magnitude of audible differences or a measure of absolute values on appropriate subjective scales.

To note is that these points will never be fully achieved but can be closely approached [38]. Point 1 was approached by always having the same set up. Point

2 was approached by using headphones instead of loudspeakers. And finally the third point was approached by using an open source MUSHRA test by Audio Labs called webMUSHRA [39].

2.8.1 The MUSHRA Test

The MUSHRA test is an international standard by the radio communications sector of the International Telecommunication Union and is called BS.1534-3 [40]. The test contains a known reference, and multiple stimuli with a hidden reference. The hidden reference is hidden in the sense that it appears to be one of the stimuli when in reality it is identical to the reference that the test participants are given. The MUSHRA test helps to quantify subjective judgements by allowing the participants to give scores of 0 to 100, where 0 is the worst and 100 the best, depending on how much they like what they hear.

Chapter 3

Implementation

In order to implement the algorithms correctly and achieve a suitable result, several aspects need to be considered before starting the implementation process. Firstly, a suitable programming language needs to be selected. This decision should be made based upon multiple factors including effectiveness, since the algorithm is aimed to run in real time, but also ease of use to minimize complex code and time taken to code the implementations. The selection of which algorithms should be implemented also needs to be considered. Quantifiable data of the algorithms need to be analyzed to be able to argue why an algorithm is promising, and should therefore be implemented, or not. In addition, if an algorithm is too complex or difficult to implement, these might be outside the scope of this thesis and could then be utilized in future work.

When both of these selections have been argued for, and made, the implementation of the algorithms could begin. These implementations will be presented and described in this chapter to show how an algorithm works. These implementations could then be tested, both by performance and subjective listening, to get quantified data on how well the algorithms enhance the audio quality.

3.1 Selection of Programming Language

To minimize the real-time latency of audio-processing the usage of a high performing language is preferred. The decision to implement in Rust was based on a preexisting library lv2-rust that implements LV2. LV2 is an audio plugin API that allows for the creation of plugins, with a minimal core specification and a design that allows for the creation of almost any feature [41]. This made real-time audio processing simple, meaning the focus could be on implementing the algorithms and methods, rather than implementing a real-time audio processor.

3.2 Selection of Algorithms

To decide which algorithms should be implemented, some metrics needed to be considered. In most cases, this metric was looking at listening tests carried out in the algorithms corresponding papers. If an algorithm was scored high, it was considered for implementation.

3.2.1 Non-Linear Device

An NLD implementation showed great promise in several papers [31], [34] achieving good bass enhancement with low distortion equating to a good over-all audio quality. Because the different NLD-implementation are very alike with the major difference only being a mathematical formula, these appeared to be quite fast to implement, allowing for more of them to be implemented.

3.2.2 Demixing and Enhancement by Non-Linear Device

This approach was highlighted in [36] and showed promising results in a listening test that measured basic audio quality and was therefore selected. Although the approach was meant to split the audio signal based on different instrumental tracks, it was deemed that with the time constraint of this project it would be enough to test it without the Music Demixing Model. In [36] it is also mentioned that a GPU is used for the music demixing model, meaning that if it were to run on the chip on the speakers it would take a longer time to process.

3.2.3 Hybrid Phase Vocoder and Non-linear Device

There were multiple papers that highlighted different hybrid PV/NLD solutions that performed well in their corresponding subjective listening tests. One of these was [30] where it was shown that a hybrid approach could improve the overall sound quality and the amount of bass compared to a regular NLD implementation. However, there was no satisfactory explanation of how the proposed method was actually achieved, and it was therefore discarded. The method that was chosen to be implemented was highlighted in [37]. It showed a significant increase in bass quality with minimal unwanted distortion compared to a regular NLD. The only issue with the selected method was a "Harmonic and Percussive Component Separation"-element which would only boost the percussive elements, such as a snare drum, instead of adding virtual bass to it. This was deemed overly complex and it was argued that it would not make a big difference in the overall quality of the algorithm and was thus was not implemented.

3.2.4 Dynamic Equalization

As outlined above, a dynamic EQ might improve the overall audio quality and was thus considered to be implemented. However, since [35] had no specific implementation mentioned, only the approach of boosting the bass depending on the headroom available was used.

3.3 Implementation of Algorithms

The plug-in can be applied to both audio files and real-time audio signals. Regardless of which application is selected, the audio signal is treated as if it was a real-time signal. For every sample the plugin receives an input sample, and outputs the modified sample. For the purpose of accurate filtering, the sample rate has to be set to the sample rate used by the speaker or audio-file. The network amplifiers used in the study were all using a sampling rate of 48 kHz.

3.3.1 Non-Linear Device

Four different NLDs were implemented, ATSR, ATSR_{var}, EXP, and NTANH. These NLDs were all implemented similarly and a diagram highlighting how it worked can be seen in Fig. A.2. The sample was first processed by a compressor,

compressing it to 6 dB headroom to guarantee that the original information contained in the signal would not be lost upon further processing. Then the sample was split into a high-frequency component and a low-frequency component. To obtain the high-frequency component, the compressed signal was simply passed through an HPF with a cut-off frequency of 200 Hz. This is also carried out for the low-frequency component but instead of an HPF, an LPF was used. The LPF used also had the cut-off frequency of 200 Hz. The low-frequency component was then processed with one of the 4 different NLDs mentioned above to generate the bass' harmonic series, before being summed together with the high-frequency component and finally being processed with a limiter to limit the output to 6 dB headroom to make sure there would be no clipping.

3.3.2 Demixing and Enhancement by Non-linear Device

The sample was passed through a compressor to normalize the input signal to have 6 dB of headroom. The sample was then delayed for 1024 samples, in order to later be used by a weighting formula. After filling the 1024 sample first in, first out(FIFO) queue, the oldest sample was then retrieved and split into two, with one being passed through an LPF with a cut-off frequency of 200 Hz and the other remaining a copy of the compressed sample. The low-passed sample was then subtracted from the copy to create a copy without a bass component, before being weighted by multiplying the low-passed sample with the following formula's result [36]:

$$\frac{\beta}{\max([x_1, x_2, x_3, ..., x_{1024}])} \tag{3.1}$$

where β is a tunable parameter to control the generation of harmonics and set to $\beta = 0.5$, and x_n is the n-th element of the FIFO-queue. The weighted sample is then processed with the following NLD, as described in the original paper [36]:

$$h(x) = \begin{cases} ATSR(x) & 0 < x \\ \tanh(2.25x) & \text{otherwise} \end{cases} \tag{3.2}$$

ATSR is the same ATSR as described in 2.7.1. The processed sample is then combined with the modified copy to be passed through a limiter to limit the output to 6 dB headroom. A block diagram highlighting the entire process can be seen in Fig. A.3.

3.3.3 Hybrid Phase Vocoder and Non-Linear Device

The input sample was compressed to leave 6 dB of headroom before being buffered for 2048 samples. The buffer enables the use of an optimized Fast Fourier Transform(FFT) algorithm called radix-2 decimation-in-time FFT. These samples were then passed through a BPF bank consisting of multiple BPF. This was done to be able to find multiple local maxima in the signal and these BPFs covered the band 40-200 Hz, as well as a HPF with a cut-off frequency of 200 Hz to keep all the non-bass frequencies. This generated one sample for each BPF and one for the HPF. These multiple band-passed samples were then converted to the frequency

domain by applying FFT. This resulted in multiple spectra consisting of multiple frequency bins. All frequency bins that did not contain the largest peak, in a spectrum, were zeroed. This resulted in multiple spectra with only one contributing frequency bin per spectrum. These spectra were then converted back to the time-domain with an inverse FFT before being passed through the following NLD [37].

$$y = HWR(x) + CLP(x) \tag{3.3}$$

with CLP(x) and HWR(x) being the following [37]:

$$CLP(x) = \begin{cases} 0.5sgn(x) & \text{if } |x| > 0.5 \\ x & \text{otherwise} \end{cases} \tag{3.4}$$

$$HWR(x) = 0.5(x + |x|) \tag{3.5}$$

Finally, the samples were summed together with the high passed sample and put through a limiter, limiting the output to a headroom of 6 dB. The entire process can be seen in Fig. A.4. On account of all samples being processed at once, resulting in 2048 samples being generated with the inverse FFT, a FIFO output queue was implemented and the calculations were only done once every 2048 samples.

3.3.4 Dynamic EQ

Showing a possible bass enhancing effect with low distortion was highlighted in [35]. By reason of it was only testing the bass enhancement and the distortion it was argued that the overall audio quality might improve compared to an unmodified signal. Since there was no specific implementation mentioned, only the approach was used. The dynamic EQ is implemented by compressing the input, creating multiple LSFs and depending on the headroom left in the signal the filter with the maximum boost possible is picked, and limiting the output. To determine the headroom of the compressed signal, 2^{13} samples are put in a buffer and the peak value is found. From these peak values the shelving filter can be picked. The thresholds were calculated as follows: Firstly the safety margin was decided to be -6 dB meaning that 0.631 is the max desired amplitude.

$$y_n = 20\log(x_n)\ \text{dB} \tag{3.6}$$

Where x_n is the n-th input sample and y_n is the n-th output sample. From this equation, different levels of gain can be calculated to thresholds for the peak values by inverting the equation.

$$x_n = 10^{\frac{y_n}{20}} \tag{3.7}$$

The calculated values are then rounded to the third decimal point.

As seen in Table 3.1, the LSF used will be decided by the amplitude of the signal sample. When extra headroom is available, the bass will be boosted to make the bass more pronounced.

However, modifying the signal through this method can create large jumps from

Gain (dB)	Threshold
+1.5	$0.501 < x_n \leq 0.562$
+3	$0.447 < x_n \leq 0.501$
+4.5	$0.398 < x_n \leq 0.447$
+6	$0.355 < x_n \leq 0.398$
+7.5	$0.316 < x_n \leq 0.355$
+9	$0.282 < x_n \leq 0.316$
+10.5	$0.251 < x_n \leq 0.282$
+12	$x_n \leq < 0.251$

Table 3.1: Table of gain depending on peak values

sample to sample, meaning one sample can be boosted +12 dB and the next +3 dB. This leads to unwanted artefacts such as crackling and distortion. To solve this problem, the level of input is not the only factor to decide what filter and boost should be used. Instead, the algorithm divides the signal into windows with 2^{13} samples. From this window, the maximum headroom is found and the gain is fetched from Table 3.1 in a tumbling window fashion. The difference is that the dynamic EQ algorithm now checks what gain was used for the former window, and if the headroom is sufficiently large the gain level can be raised by one increment. Or, if it is smaller than for the previous window, the algorithm can either choose one or two levels of lower gain depending on how much less headroom there is.

A simplified example of a signal can be seen in Fig. 3.1. If the initial boost is set to an arbitrary but low level, lets say +3 dB as the initial point, then the max amplitude from window 1, being 0.3, will fetch a boost of +9 dB from the table. To avoid these big jumps the level of boost will instead increase by one level in the table becoming +4.5 dB. The same happens again with the amplitude of window 2, being 0.3, and the boost level becomes +6 dB. For window 3 the amplitude is 0.6 leaving little headroom, now the level will go down one step to +4.5 dB to avoid clipping. For the 4th and 5th window, both will have the required headroom to further boost the signal resulting in a gain of 6 dB and 7.5 dB respectively. This happens for every window creating a smoother dynamic EQ, minimizing the level of clipping and crackling.

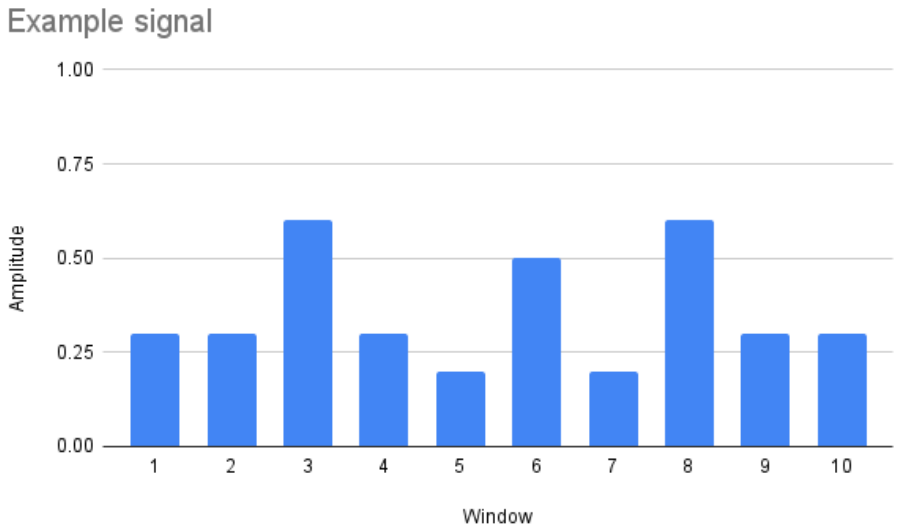

Figure 3.1: A simplified example of a signal broken down in to windows

Chapter 4

Test Setup, Results and Evaluation

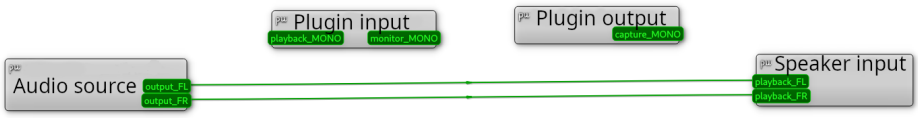

Figure 4.1: PipeWire before connecting the audio source to the algorithm.

Figure 4.2: PipeWire after connecting the audio source to the algorithm.

With the implementation done the algorithms need to be evaluated. This was achieved with the help of a performance test and a listening test. The purpose of the performance test was to see if an algorithm is actually able to run in real-time on a speaker, and the listening test to determine which, if any, actually enhances the perceived audio quality.

4.1 Performance Test Setup

The performance tests were performed by testing all the selected algorithms in real time on the speakers. In order to measure the CPU usage of the algorithms, the plugin had to be put on the speaker chip. By using PipeWire, a free open source multimedia handler that allows for the rerouting of multimedia pipelines, it was possible to route the audio input through the LV2 plugin, i.e. the algorithm in use, and out to the speaker. The routing in PipeWire is illustrated in Fig. 4.1 and 4.2 where the first figure shows the routing before connecting the plugin and the second shows the plugin connected. This means that the algorithm could now be run in real time and CPU measurements could be obtained. The measurements were obtained by connecting to the speaker by secure shell (SSH). Running the top command [42], which shows how much of the CPU is used by each process, the usage of the plugin could be read and recorded. Examples of the speakers and amplifier tested can be seen in Fig. 4.3 and 4.4.

4.2 Listening Test Setup

As previously mentioned, there was a need to get quantifiable data on how well the perceived audio quality of the algorithms sounded. This was done by implementing the MUSHRA test in a form of webMushra [39]. Before the tests could

(a) Speaker with less powerful chip. **(b)** Speaker with more powerful chip.

Figure 4.3: The two speakers tested.

commence, the audio clips had to be prepared. For the sake of simplicity and compatibility with the webMUSHRA interface, the clips had to be preprocessed even though the algorithms are intended to run in real time. Clips from different genres were processed, and synced and cut with the help of Audacity to 12 seconds since the webMUSHRA service only supports clips of 12 seconds or shorter. For the speaking clips, the clips were run through the algorithms before the test and also processed in the same way. The test setup was the webMUSHRA interface, using a computer seen in Fig. 4.5. The loudspeakers that were used, were in the form of the Beyerdynamic DT 770 PRO 250 ohm electrodynamic headphones as can be seen in Fig. 4.6.

In this test, the participants were informed to consider the reference signal to be 50 on a scale of 0 to 100. Then the participants would rate the stimuli, in this test the different algorithms, as better or worse than the reference signal. The participants consisted of audio engineers and other employees of the partnered company. When arriving, they were introduced to the MUSHRA test, where some participants were already familiar with the procedure. The participants were instructed to subjectively rate each clip on the scale of 0 to 100, based on how they thought the audio quality was, while taking into account any distortion, clipping, or other disturbance they might hear. They were also told to consider the reference as 50. If they thought an algorithm sounded better they should give a score higher than 50 and the if opposite was true they should give it lower than 50. However, if they thought it sounded equally good, they should give it 50. Then the test would commence, starting with a page where the participant could familiarize themselves with the webMUSHRA tool. After they felt comfortable with the tool, they would continue with the tests. Firstly, the music clips were presented to the participants in a random order, and afterwards the speaking clips were presented, also in a random order. When the participants were finished they submitted the answers.

Figure 4.4: The network amplifier tested.

Figure 4.5: Example of a webMUSHRA page.

(a) Front view. **(b)** Side view.

Figure 4.6: The beyerdynamic DT 770 PRO 250 ohm used for the MUSHRA listening test.

4.3 Analysis

After the listening and performance tests, the results were analyzed to find trends in the data. The results were collected in tables as well as plotted with the help of python and matplotlib, a plotting package for python. These results are presented later in this chapter together with explanations of the figures and tables. These were then explored on a deeper level and put into context in the discussion.

4.4 Selection of Algorithms

To be able to transition from all the algorithms presented earlier in the report, to a more suitable amount of algorithms for the participants to rate, they were tested with an initial listening test before commencing the main listening tests of this project. This was done in order to select a couple of algorithms that were promising. The quality of an algorithm could also be observed when inspecting the sound wave for clipping and jumping using Audacity, as can be seen in Fig. 4.7 and 4.8. This was made quite easy since there were only a couple of algorithms that did not explicitly sound bad. However, if they were better than the unprocessed was yet to be determined by the test participants. All of the basic NLDs described

in the previous section had no major issues with either clipping or jumping. ATSR had arguably the best scores in the case study when comparing the different NLDs and was therefore the only basic NLD that would proceed to the listening test. This was done not only to diversify the methods used, but also as mentioned above, to have a more suitable amount of algorithms for the test subject to rate. The chosen algorithms put into the listening test were:

- ATSR
- Demixing and enhancement by NLD (Demix)
- Dynamic EQ (DEQ)

Figure 4.7: Example of a clipping signal, leading to a loss of information.

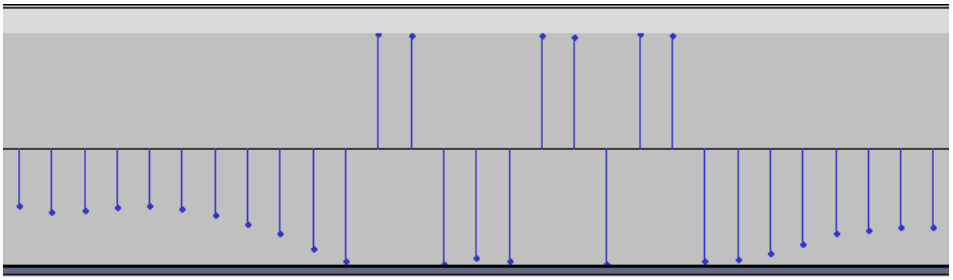

Figure 4.8: Example of a jumping signal, creating unwanted artefacts.

4.5 Performance Readings

The performance readings were measured on devices with three different CPUs. These were all measured while streaming music over a network connection. The tables (Table 4.1, 4.2, 4.3) show how much of the CPU is used. In addition, they also show what window type, and window length that is potentially used for each of the algorithms.

Algorithm	CPU usage	Window type	Window size
None	∼1%	N/A	N/A
ATSR	∼11%	N/A	N/A
Demix	∼87%	Sliding	1024
Dynamic EQ	∼5%	Tumbling	8192

Table 4.1: Performance for different algorithms on a NXP i.MX 6ULL chip with 256 MB of RAM

Algorithm	CPU usage	Window type	Window size
None	∼1%	N/A	N/A
ATSR	∼8%	N/A	N/A
Demix	∼62%	Sliding	1024
Dynamic EQ	∼3%	Tumbling	8192

Table 4.2: Performance for different algorithms on a NXP i.Mx 6SoloX chip with 512 MB of RAM

Algorithm	CPU usage	Window type	Window size
None	<1%	N/A	N/A
ATSR	∼1%	N/A	N/A
Demix	∼12%	Sliding	1024
Dynamic EQ	∼1%	Tumbling	8192

Table 4.3: Performance for different algorithms on a NXP i.MX 8M Nano chip with 1024 MB of RAM

4.6 Listening Test Results

The data from the listening test were collected in a comma separated value (CSV) file and was plotted using matplotlib in python. For both the music and speaking clips, the plot consists of box plots describing the scores achieved for each algorithm and music clip, while the bar chart shows the average score of each algorithm applied on each music clip. On the X-axis the clips are presented one by one and on the Y-axis the score from 0 to 100 is shown.

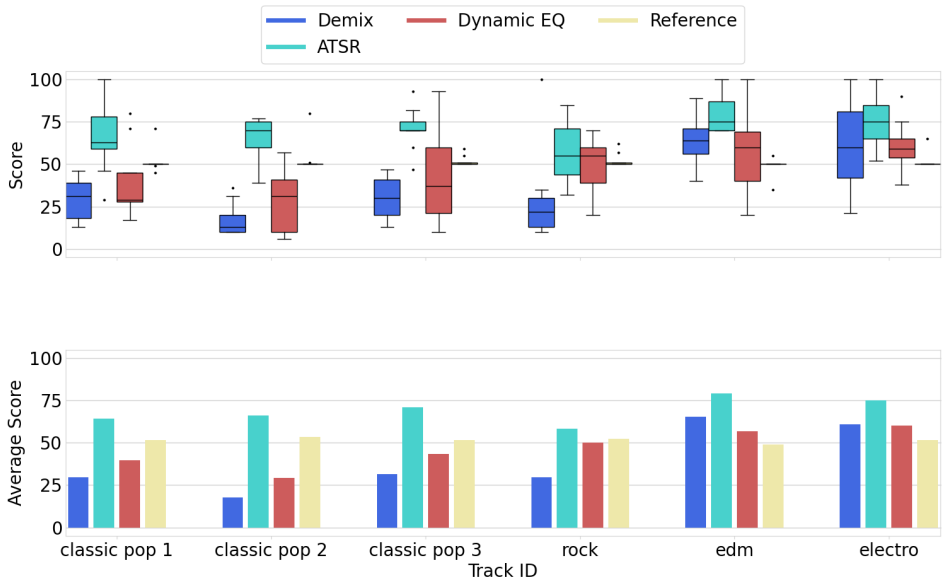

Figure 4.9: Plot of the scores in a box plot, together with a bar chart of the average scores for each algorithm and music genre.

4.7 Reflections

In this section, reflections of the three algorithms will be presented and completed with a comparison between them. The results from the performance readings and listening tests will be analyzed comprehensively to get a clear view of which algorithm could be considered for further use and development.

4.7.1 Basic Non-Linear Device (ATSR)

When observing Fig. 4.9 it is easy to see that the ATSR algorithm scores higher for each of the music clips . It is consistently scored with the highest median and mean score across the different genres, with rock being the only genre not improved drastically from the algorithm being applied. However, even with the score being lower it is still an improvement on the reference and is also scored higher than the

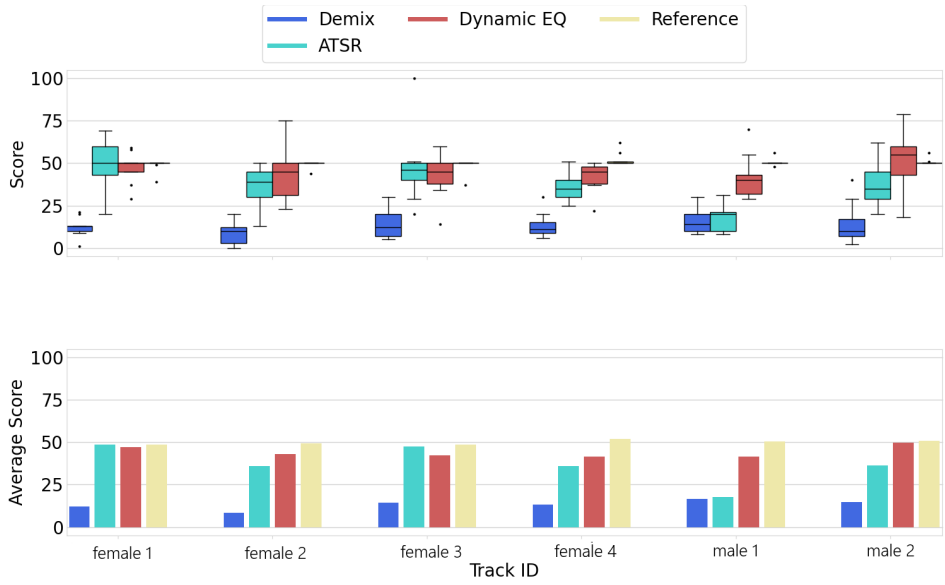

Figure 4.10: Plot of the scores in a box plot, together with a bar chart of the average scores for each algorithm and speaking clip.

other algorithms.

On the other hand, looking at Fig. 4.10 instead, it can be noted that the algorithm is scored equal to the reference for two clips, (female 1 and female 3), and lower for the other four clips. This could be due to the speaking clips containing more noise than the music clips. Speaking also has a different frequency range than that of music, leading to different behaviour from the algorithm.

It could also be argued, from Tables 4.1, 4.2 and 4.3, that ATSR as an algorithm does not have high demands on a CPU since it only demands roughly 11% for the weakest of the three measured CPU's.

With all this in mind, the ATSR algorithm could be considered a reasonable approach to enhance the perceived audio for music streaming.

4.7.2 Demixing and Enhancement by Non-linear Device (Demix)

From Fig. 4.9, it is also possible to deduct that the demix algorithm scored worse than compared to both the other algorithms and the unaltered reference signal for classic pop and rock. However, the algorithm scores high points for the edm and electro clips. The demix algorithm scores the lowest out of all algorithms when it comes to audio clips where someone is speaking, and offers no improvement to the reference signal as can be seen in Fig. 4.10. The worse performance for speaking and the non-synthesised songs, could be attributed to higher noise levels that

might be present in the original signal that the algorithm is not able to handle correctly. This possible source of error is strengthened by the scores given for the edm and electro clips, since these songs are modern and synthesized, eliminating many sources of noise introduction from guitars, cables etc.

From Tables 4.1, 4.2 and 4.3, it is clear that the demix algorithm is inefficient and has higher demands on the CPU's power, only being a reasonable consideration for powerful CPU's. This is in large part due to the algorithm having a sliding window creating queues of 1024 samples for each sample being processed and calculated.

4.7.3 Dynamic EQ (DEQ)

For the dynamic EQ algorithm it can be analyzed from Fig. 4.9 how the results vary from different genres with a better score for synthesized music. Because it is boosting from a low frequency it can potentially boost noise and not only the real signal. This would explain why it scored much worse on the music with real instruments compared to the music that is electronic. Because the algorithm will always boost the signal, parts of the songs where the music is quiet will be able to boost noise even more leading to a worse audio quality. However, when observing Fig. 4.10 this is not reflected. For the announcements, the dynamic EQ scores the highest points of the algorithms while being close to the reference signal. The observation that it does not score as bad as the other two algorithms does not mean it should be used as an algorithm for announcements. Rather, it could indicate that the dynamic EQ does not boost anything for these clips. This could be due to the audio track being normalized in a way that does not leave any extra headroom for the algorithm to boost.

It is clear from the data in Tables 4.1, 4.2 and 4.3, that the dynamic EQ is a very efficient algorithm that could be considered for even weaker CPU's than the ones measured.

4.7.4 Summary of Algorithms

From Fig. 4.10 it can easily be seen that there was no algorithm that showed any significant improvement for clips of people speaking or announcing. This could be due to multiple factors, such as distortion or unwanted artefacts already being present in the original signal, as well as a different frequency spectrum than that of music. This is however no issue, since the algorithms can be turned off for scheduled announcements for any network loudspeaker running a PipeWire style software.

The results from the listening tests for music clips, seen in Fig. 4.9, show that certain algorithms does enhance the overall perceived audio quality on certain genres of music. ATSR shows a median and average score being higher than that of the reference as well as compared to the other algorithms. This would indicate that one of the most lightweight algorithms, as can be seen in Tables 4.1, 4.2 and 4.3,

also has the best overall performance according to the scores from the listening tests.

Comparing all of the results from the listening tests as well as the performance of the algorithm, it is clear that the ATSR algorithm should be considered for implementation and future development.

Chapter 5
Conclusion

The improvement of audio quality is important due to the vast expansion of the usage of loudspeakers which has led to demands on cost, size, connectivity, and other design choices, all leading to different loudspeaker characteristics. A common characteristic for all smaller loudspeakers is the impaired ability to reproduce low-frequency signals. In this master's thesis, the primary aim is to improve the perceived audio quality of a loudspeaker, specifically a network speaker powered by PoE, with the help of digital signal processing. The aim of the thesis also includes researching how the digital signal processing can be performed under various conditions and how the processing can be adjusted to handle sudden signal changes. The result of the literature study conclude that a VBE algorithm, as well as a dynamic EQ algorithm, is the best alternative to enhance the audio quality. The implemented dynamic EQ possesses a higher resilience to sudden changes in signal level, reducing the number of artefacts in a processed signal. From the various algorithms, the VBE algorithm, ATSR, was determined to give the best result by both scoring the highest of the algorithms tested in a listening test and showing a good performance in a performance test. ATSR also showed that different music clips with varying volume could be handled and produce an improvement compared to a reference. None of the algorithms showed any significant improvement when applied to announcements. However, the algorithms can be turned off for scheduled announcements. The simplicity of the algorithm together with the improvement it offers, leads to it being a good option for any network speaker. Therefore, the implementation of the ATSR algorithm is recommended for immediate use.

5.1 Future Research

As outlined above, the noise seems to be an issue for all of the algorithms and therefore future research could investigate trying to reduce noise on the input signal to enhance the perceived audio quality. Reducing the noise of the signal could possibly improve the dynamic EQ algorithm by making it only boost the actual audio signal and not the noise around it. An improved dynamic EQ could in turn be combined with an NLD for a combination of a bass-boost and a virtual-bass algorithm as done in [35]. In addition to improving the algorithms by noise reduction, experimenting with the parameter setup of the ATSR algorithm, as described in [43], could further improve audio enhancement by fine tuning the algorithm. Finally the demixing algorithm could be improved by implementing the Music Demixing Model that was mentioned in the original article [36] instead of using a simple LPF. This could improve the demixing algorithm by boosting the instruments separately, gaining improvement in quality for each instrument and not only for the overall audio.

The listening tests in this thesis only covered electrodynamic headphones, which in future work could be extended to test other loudspeaker types. If the listening tests were expanded to test multiple different loudspeaker types, they could show if the found improvement is consistent across other types or is limited to only electrodynamic loudspeakers.

Bibliography

[1] C. Gazzola, V. Zega, F. Cerini, S. Adorno, and A. Corigliano, "On the design and modeling of a full-range piezoelectric mems loudspeaker for in-ear applications," *Journal of Microelectromechanical Systems*, vol. 32, no. 6, pp. 626–637, 2023.

[2] M.-C. Cheng, W.-S. Huang, and S. R.-S. Huang, "A silicon microspeaker for hearing instruments," *Journal of Micromechanics and Microengineering*, vol. 14, p. 859, may 2004.

[3] I. Shahosseini, E. Lefeuvre, E. Martincic, M. Woytasik, J. Moulin, S. Megherbi, R. Ravaud, and G. Lemarquand, "Microstructured silicon membrane with soft suspension beams for a high performance mems microspeaker.," *Microsystem Technologies*, vol. 18, no. 11, pp. 1791 – 1799, 2012.

[4] W. H. Watkins, *Loudspeaker Physics and Forced Vibration*. Springer Cham, 2022.

[5] C.-M. Lee, J.-H. Kwon, K.-S. Kim, J.-H. Park, and S.-M. Hwang, "Design and analysis of microspeakers to improve sound characteristics in a low frequency range," *IEEE Transactions on Magnetics*, vol. 46, no. 6, pp. 2048–2051, 2010.

[6] E. Sturtzer, I. Shahosseini, G. Pillonnet, E. Lefeuvre, and G. Lemarquand, "High fidelity microelectromechanical system electrodynamic micro-speaker characterization," *Journal of Applied Physics*, vol. 113, p. 214905, 06 2013.

[7] R. Larson, "The electrostatic loudspeaker–an objective evaluation," *IRE Transactions on Audio*, vol. AU-4, no. 2, pp. 32–36, 1956.

[8] H. Conrad, H. Schenk, B. Kaiser, S. Langa, M. Gaudet, K. Schimmanz, M. Stolz, and M. Lenz, "A small-gap electrostatic micro-actuator for large deflections.," *Nature Communications*, vol. 6, 2015.

[9] B. Kaiser, F. Wall, J. Monsalve, S. Langa, M. Stolz, A. Melnikov, D. Schuffenhauer, H. Schenk, H. Schenk, L. Ehrig, and H. Conrad, "The push-pull principle: an electrostatic actuator concept for low distortion acoustic transducers.," *Microsystems and Nanoengineering*, vol. 8, no. 1, 2022.

[10] T. Cheng and J. Li, *Piezoelectric Actuators*. IntechOpen, 2022.

[11] J. W. Suk, K. Kirk, Y. Hao, N. A. Hall, and R. S. Ruoff, "Thermoacoustic sound generation from monolayer graphene for transparent and flexible sound sources," *Advanced Materials*, vol. 24, no. 47, pp. 6342–6347, 2012.

[12] L. Xiao, Z. Chen, C. Feng, L. Liu, Y. Wang, L. Qian, Y. Zhang, Q. Li, K. Jiang, S. Fan, and Z.-Q. Bai, "Flexible, stretchable, transparent carbon nanotube thin film loudspeakers.," *Nano Letters*, vol. 8, no. 12, pp. 4539–4545 – 4545, 2008.

[13] W. Fei, J. Zhou, and W. Guo, "Low-voltage driven graphene foam thermoacoustic speaker," *Small*, vol. 11, no. 19, pp. 2252–2256, 2015.

[14] H. Tian, T.-L. Ren, D. Xie, Y.-F. Wang, C.-J. Zhou, T.-T. Feng, D. Fu, Y. Yang, P.-G. Peng, L.-G. Wang, and L.-T. Liu, "Graphene-on-paper sound source devices.," *ACS Nano*, vol. 5, no. 6, pp. 4878–4885 – 4885, 2011.

[15] Axis Communications, *AXIS C1410 Network Mini Speaker*, 2023. Accessed 15 May. 2024.

[16] Axis Communications, *AXIS C8210 Network Audio Amplifier*, 2023. Accessed 12 Feb. 2024.

[17] Axis Communications, *AXIS C1111-E Network Cabinet Speaker*, 2024. Accessed 15 May. 2024.

[18] J. G. Proakis and D. K. Manolakis, *Digital Signal Processing*, ch. 1, 5. PEARSON, 2014.

[19] "IEEE standard for information technology - telecommunications and information exchange between systems - local and metropolitan area networks - specific requirements - part 3: Carrier sense multiple access with collision detection (csma/cd) access method and physical layer specifications - data terminal equipment (dte) power via media dependent interface (mdi)," *IEEE Std 802.3af-2003 (Amendment to IEEE Std 802.3-2002, including IEEE Std 802.3ae-2002)*, pp. 1–133, 2003.

[20] "Ieee standard for ethernet amendment 2: Physical layer and management parameters for power over ethernet over 4 pairs," *IEEE Std 802.3bt-2018 (Amendment to IEEE Std 802.3-2018 as amended by IEEE Std 802.3cb-2018)*, pp. 1–291, 2019.

[21] L. Khichadi and K. Nagamani, "Performance evaluation of power over ethernet in an ethernet switch," in *2019 International Conference on Communication and Electronics Systems (ICCES)*, pp. 1091–1095, 2019.

[22] D. M. Howard and J. A. S. Angus, *Acoustics and psychoacoustics. [Elektronisk resurs].* Focal, 2009.

[23] L. Wang, W. S. Gan, Y. K. Chong, and S. Kuo, "A novel approach to bass enhancement in automobile cabin," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 3, pp. III–III, May 2006.

[24] E. Perez-Gonzalez and J. Reiss, "Automatic equalization of multichannel audio using cross-adaptive methods," in *Audio Engineering Society Convention 127*, Audio Engineering Society, 2009.

[25] F. de Dieuleveult, "4 - amplifier output stage," in *Amplifiers and Oscillators Optimization by Simulation* (F. de Dieuleveult, ed.), pp. 139–160, Elsevier, 2018.

[26] P. M. Todd and D. G. Loy, *Music and connectionism*. Mit Press, 1991.

[27] A. Seebeck, "Ueber die erzeugung von tönen durch getrennte eindrücke, mit beziehung auf die definition des tones," *Annalen der Physik*, vol. 139, no. 11, pp. 368–380, 1844.

[28] Z. Průša and N. Holighaus, "Phase vocoder done right," in *2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 976–980, 2017.

[29] N. Akaishi, K. Yatabe, and Y. Oikawa, "Improving phase-vocoder-based time stretching by time-directional spectrogram squeezing," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

[30] S. C. Pulikottil, "Virtual bass system by exploiting the rhythmic contents in music," in *2015 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*, pp. 1–5, 2015.

[31] R. Giampiccolo, A. Bernardini, and A. Sarti, "A time-domain virtual bass enhancement circuital model for real-time music applications," in *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–5, Sep. 2022.

[32] E. Larsen and R. Aarts, "Reproducing low-pitched signals through small loudspeakers," *Journal of the Audio Engineering Society. Audio Engineering Society*, vol. 50, 01 2012.

[33] S. Zhang, L. Xie, Z.-H. Fu, and Y. Yuan, "A hybrid virtual bass system with improved phase vocoder and high efficiency," in *The 9th International Symposium on Chinese Spoken Language Processing*, pp. 401–405, Sep. 2014.

[34] N. Oo and M. Hawksford, "Perceptually-motivated objective grading of nonlinear processing in virtual bass systems," *Journal of the Audio Engineering Society*, 11 2011.

[35] B. Pueo, G. Ramos, and J. J. Lopez, "Strategies for bass enhancement in multiactuator panels for wave field synthesis," *Applied Acoustics*, vol. 71, no. 8, pp. 722–730, 2010.

[36] R. Giampiccolo, A. I. Mezza, A. Bernardini, and A. Sarti, "Virtual bass enhancement via music demixing," *IEEE Signal Processing Letters*, vol. 30, pp. 908–912, 2023.

[37] T. Lee, S. Lee, Y.-c. Park, and D. H. Youn, "Virtual bass system based on a multiband harmonic generation," in *2013 IEEE International Conference on Consumer Electronics (ICCE)*, pp. 399–400, 2013.

[38] F. E. Toole, "Listening tests - turning opinion into fact," vol. 30 of *JAES*, (Los Angeles), pp. 431–445, Audio Engineering Society, June 1982. aff. National Research Council, Ottawa, Ont. K1A OR6, Canada.

[39] M. Schoeffler *et al.*, "webmushra — a comprehensive framework for web-based listening tests," *Journal of Open Research Software*, vol. 6, no. 1, p. 8, 2018.

[40] R. Sector, *Method for the subjective assessment of intermediate quality level of audio systems*. International Telecommunication Union, 2015.

[41] LV2, "Why lv2?." lv2plug.in. Accessed: June. 14, 2024. [Online]. Available: https://lv2plug.in/pages/why-lv2.html.

[42] M. Kerrisk, "top(1) - linux manual page." man7.org. Accessed: May. 22, 2024. [Online]. Available: https://man7.org/linux/man-pages/man1/top.1.html.

[43] N. Oo, W.-S. Gan, and W.-T. Lim, "Generalized harmonic analysis of arctangent square root (atsr) nonlinear device for virtual bass system," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 301–304, 2010.

Appendix A

Block Diagrams

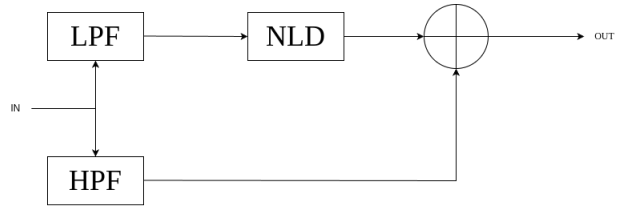

Figure A.1: A block diagram showing a simple NLD implementation.

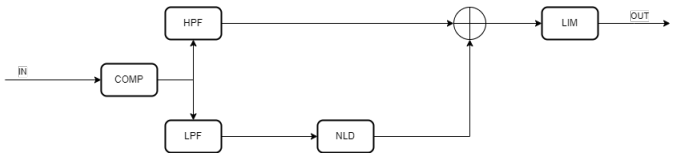

Figure A.2: A block diagram showing an NLD implementation with a compressor and limiter.

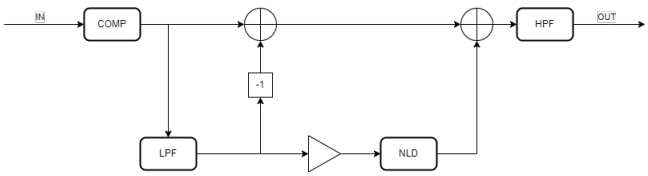

Figure A.3: A block diagram showing the demixing algorithm.

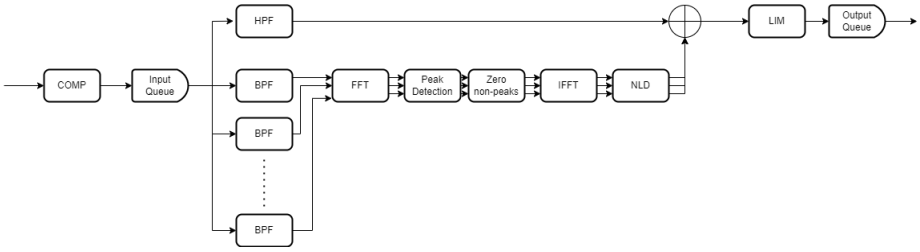

Figure A.4: A block diagram showing the hybrid PV and NLD algorithm.

Printed by Tryckeriet i E-huset, Lund 2024

LUND
UNIVERSITY

Series of Master´s theses
Department of Electrical and Information Technology
LU/LTH-EIT 2024-999
http://www.eit.lth.se